Asymptotically Optimal Method for Manifold Estimation Problem

POSTER
Room
SUN
18:45-20:00

ALEXANDER KULESHOV*, ALEXANDER BERNSTEIN†,*,‡, YURY YANOVICH*

*Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia,
†Institute for System Analysis, Russian Academy of Sciences, Moscow, Russia
‡email: a.bernstein@mail.ru

Let $\mathbf{X}$ be an unknown nonlinear smooth q-dimensional Data manifold (D-manifold) embedded in a p-dimensional space ($p > q$) covered by a single coordinate chart. It is assumed that the manifold's condition number is positive so $\mathbf{X}$ has no self-intersections. Let $\mathbf{X}_n = \{X_1, X_2, \ldots, X_n\} \subset \mathbf{X} \subset \mathbf{R}^p$ be a sample randomly selected from the D-manifold $\mathbf{X}$ independently of each other according to an unknown probability measure on $\mathbf{X}$ with strictly positive density. The Manifold Estimation problem (ME) is to construct sample-based q-dimensional manifold (Estimated Data Manifold, ED-manifold) $\mathbf{X}_\theta \subset \mathbf{R}^p$ covered by a single coordinate chart which is close to the D-manifold $\mathbf{X}$.

The problem solution $\theta = (h, g)$ consists of two sample-based interrelated mappings: an Embedding mapping $h\colon\ \mathbf{X}_h \to \mathbf{R}^q$ defined on the domain $\mathbf{X}_h \supseteq \mathbf{X}$, and a Reconstruction mapping $g\colon\ \mathbf{Y}_g \subset \mathbf{R}^q \to \mathbf{R}^p$ defined on the domain $\mathbf{Y}_g \supseteq h(\mathbf{X}_h) \supset h(\mathbf{X})$. The solution θ determines the ED-manifold $\mathbf{X}_\theta = \{X = g(y) \in \mathbf{R}^p\colon\ y \in \mathbf{Y}_\theta \subset \mathbf{R}^q\}$ embedded in $\mathbf{R}^p$-dimensional space and covered by a single coordinate chart g defined on space $\mathbf{Y}_\theta = h(\mathbf{X})$ and must ensuring the accurately reconstruction $\mathbf{X}_\theta \approx \mathbf{X}$ of the D-manifold (Manifold proximity property which implies the approximate equalities $X \approx g(h(X))$ for all $X \in \mathbf{X}$).

In [Bernstein & Kuleshov, 2012] an amplification of the ME called the Tangent Bundle Manifold Learning (TBML) is proposed. The TBML problem is to estimate a tangent bundle $\{(X, L(X)), X \in \mathbf{X}\}$ consisting of the points X from the D-manifold $\mathbf{X}$ and the tangent spaces $L(X)$ at these points. The TBML solution (θ, G) includes additionally the sample-based $p \times q$ matrices $G(y),\ y \in \mathbf{Y}_g$, such that the linear space $\mathrm{Span}(G(h(X)))$ accurately reconstructs the tangent space $L(X)$ for all $X \in \mathbf{X}$. A new geometrically motivated algorithm called Grassmann&Stiefel Eigenmaps (GSE) that solves the TBML problem and gives a new solution for the ME problem is proposed also in this paper.

We present also some asymptotic properties of the GSE. In particular, it is proven that under the appropriate chosen GSE parameters there exists a number C_{GSE} such that with high probability (whp) the inequalities

$$\|X - g(h(X))\| \le C_{GSE} \times n^{-\frac{2}{q+2}}$$

for all $X \in \mathbf{X}$ hold true. Here the phrase "an event occurs whp" means that the event occurs with probability at least $(1 - c_\alpha / n^\alpha)$ for any $\alpha > 1$ and c_α depends only on α.

The achieved convergence rate $O(n^{-\frac{2}{q+2}})$ coincides with a minimax lower bound for Hausdorff distance between the manifold and its estimator in the Manifold estimation problem obtained in [Genovese et al. (2012)] under close assumptions. So, GSE has optimal rate of convergence.

Acknowledgment. This work is partially supported byLaboratory for Structural Methods of Data Analysis in Predictive Modeling, MIPT, RF government grant, ag. 11.G34.31.0073.

Bibliography

[Bernstein & Kuleshov, 2012] Bernstein A.V., Kuleshov A.P., 2012: Tangent Bundle Manifold Learning via Grassmann & Stiefel Eigenmaps. arXiv: 1212.6031v1 [cs.LG].

[Genovese et al. (2012)] Genovese Christopher R., Perone-Pacifico Marco, Verdinelli Isabella, Wasserman Larry, 2012: Minimax Manifold Estimation. *JMLR*, **13**, 1263 - 1291.